Themed Issue: Population Pharmacokinetics - In Memory of Lewis Sheiner
Guest Editors - Peter Bonate and Diane Mould

It's Time

Submitted: June 17, 2005; Accepted: June 17, 2005; Published: October 24, 2005

Brian P. Smith[1]

[1]Eli Lilly and Company, Lilly Corporate Center, Drop Code 2233, Indianapolis, IN 46285

ABSTRACT

Statistical inference involves taking the results of models and knowledge about probability to make decisions about the relationship in question. This commentary explains the usefulness of statistical inference to the drug development process, as well as some common pitfalls. It also examines reasons why statistical inference does not seem to be fully integrated into pharmacometric modeling. An example is shown that demonstrates the inferential advantages of mechanistic models. Both statisticians and pharmacometricians ought to take note of these advantages and integrate their efforts in order to maximize the decision-making potential of clinical research.

KEYWORDS: mechanistic models, statistical inference

INTRODUCTION

I have worked several years attempting to make statistics, both as a tool and a profession, a more integrated partner in clinical pharmacology research. Lewis Sheiner spent an illustrious career making clinical pharmacology a more quantitative science. Although different goals, there is enough similarity in them that Lewis and I would often discuss at meetings quantitative issues related to clinical pharmacology. The last time that I saw Lewis was at the 2004 annual American Society of Clinical Pharmacology and Therapeutics (ASCPT) meeting in Miami. I was expressing some dissatisfaction with the progress we had made and Lewis told me, basically, that he was passing on the torch to people like me to push forward the quantification of clinical pharmacology science. With his passing just a few months later, I was left with both an eerie feeling and awesome responsibility.

THE ISSUE

Clinical pharmacology is a quantitative discipline. First, it is a field that generates data. Any field where one collects data in order to understand the way things work is quantitative in nature. Yet, to add to this, it is rich in examples of being able to describe relationships through the use of mathematical models. This special issue alone presents many examples of describing relationships with rather complex nonlinear models.

Statistics is the science of quantification. Modern statistical science has its roots going back to the 17th century, but it really began its heyday at the turn of the 20th century.[1] Because of the lack of computing power many of the early problems were solved using linear models. The mathematics for developing parameter estimates under a linear model is relatively easy. Even though sometimes the use of linear models was a crude approximation of the truth, they allowed for insight, which up to that point was not possible. If the problem was sufficiently nonlinear to make the use of a linear relationship suspect, one could often answer important questions by not assuming any prespecified relationship between the dependent and independent variable by using analysis of variance (ANOVA) techniques, which are mathematically in the class of linear models. Statisticians, however, played an essential role in developing software that could examine nonlinear models. Yet, in practice, the default is often linear models (linear regression or ANOVA techniques). The question is why? Possibly part of the answer is due to the mathematical simplicity and elegance of linear models as well as to lack of training using more advanced statistical analysis methods. There is also probably a component of reliance on what has worked in the past. But probably the most compelling reason is that to use a nonlinear model, it must be first specified. Of course, one could always look at one's data and try to empirically specify a nonlinear model that appears to fit the data well. Although still in the mathematical class of linear models, the use of a quadratic term in a linear regression is one such way this is done. Yet, if there were some sound scientific principle that was driving the nonlinear relationship, it seems obvious that this would be preferred to any empirical relationship. Thus, to specify a mechanistically driven nonlinear model one must have an understanding of the scientific mechanisms that are driving the nonlinearity.

Paradoxically, the strength of statistics is also its weakness. Any science that collects and analyzes data uses statistics to understand the relationships between variables. Unfortunately, it would be impossible for organizations that train statisticians to also give them the scientific training necessary to apply nonlinear models to all of the disciplines that

Corresponding Author: Brian P. Smith, Eli Lilly and Company, Lilly Corporate Center, Drop Code 2233, Indianapolis, IN 46285. Tel: (317) 277-7315; Fax: (317) 277-3220; E-mail: b.smith@lilly.com

use statistics. Yet, the importance of statistics does not rely on being able to model data. Its importance comes from being able to take the results of models and knowledge about probability to make decisions about the relationship in question. Let's call this statistical inference.

Pharmacometrics is a relatively new discipline. It is made up mostly of individuals trained in pharmacokinetics, although there are successful statisticians who call themselves pharmacometricians as well. Anyone who has looked at a concentration-time plot (the basic plot of pharmacokinetics) recognizes immediately that pharmacokinetic relationships are nonlinear. Through the use of polyexponential equations, concentration data can be modeled quite well over time. Thus, the pharmacometrician quite naturally sees the nonlinearity of biological process. Immediately then the pharmacometrician also sees the nonlinearity associated with drug response. Through the use of similar techniques used in describing the nonlinear relationship of pharmacokinetics, they are able to develop models that describe the nonlinear relationships associated with pharmacodynamic response. Unfortunately, statistical inference is not then always pursued. Thus, a good description of the relationship has been developed, but decision making, rather than being probability-based, is often left to gut instinct. The question is why? First, the typical pharmacometrician is not trained in probability. Thus, there may be a certain amount of ignorance of the usefulness of statistical inference. In addition, it must be recognized that nonstatisticians and statisticians alike so often misinterpret statistical inference that there are many examples of it not providing insight but instead providing nonsense. Much of this revolves around the general belief that a *P* value less than .05 implies that an effect is present, whereas a *P* value greater than .05 means that no effect is present. Under some circumstances, this sort of logic can be useful, but without understanding the basic meaning of a *P* value one can find themselves making illogical decisions.

The concept of a *P* value is relatively simple, but often misunderstood. First, define the null hypothesis, ie, $\mu_t - \mu_p = 0$, as the hypothesis of no difference in means between groups. This is usually what one wants to prove is false. After one has one's data, the following thought process takes place: (1) The differences between the means, $\bar{x}_t - \bar{x}_p$ in this experiment is Δ. (2) For sake of argument, assume that $\mu_t - \mu_p = 0$. (3) With this assumption, how likely would it be that an experiment would produce an $\bar{x}_t - \bar{x}_p \geq \Delta$. (4) If this chance were relatively small, one would conclude that it was unlikely that the null-hypothesis were true. The reliance of .05 as "statistically significant" can cause many problems.[2,3] One of those problems involves the following. First, one may have too little data to disprove the null hypothesis at this significance level. For instance, it may be that given one's data and the situation at hand that a *P* value = .2 would be sufficiently small for one to conclude that the null-hypothesis is unreasonable. On the other hand, in a data-rich environment one may prefer that before concluding that the null hypothesis is unreasonable, the *P* value be less than .005. One's decision-making process is dependent upon things like the importance of making the right decision.

With standard statistical models it is common to generate many *P* values for the same response. For instance, in a repeated measures design with 6 doses (1 being placebo) and each patient is assessed 10 times over some period of time, one has 50 comparisons to placebo. If all 50 null hypotheses were true, it would obviously be very likely that some *P* values would be less than .05; an effect termed multiplicity.[4] Thus, if doing this sort of analysis, one would expect to see some sort of pattern in order to form the belief that something was really happening. Clever modeling, however, can alleviate this problem of multiplicity. If one understood mechanistically both the relationship over time and the relationship with dose, very likely one could reduce the problem to whether 1 parameter is different from 0.

The *P* value in statistics is probably over-used in statistical inference and says nothing about the magnitude of the difference between treatments. In other words, a *P* value of .0001 does not imply that the magnitude of difference between treatments is larger than when the *P* value is .01. Another classical tool is the confidence interval (CI) approach, which is advocated by many medical journals to replace the *P* value approach, because the former does assess the magnitude and precision of the difference between treatments. Suppose one could run an experiment over and over again. A 95% CI is generated in such a way that for 95 out of 100 experiments, the CI would contain the true value. The unsettling fact is that a CI will not contain the true value 5 out of 100 times. When one carries out the experiment, one does not know whether one's CI contains the truth or not. All one knows is that there was a 95% chance that it would. As with the *P* value, the level of a CI should be dependent upon the situation. Once this level is chosen, it is practical to think of the resulting CI as where the truth lies, even though this in not precisely what a CI means. By using both a CI and a *P* value one has a more robust means of doing statistical inference. It is usually not the goal to be different from placebo, but to be clinically different from placebo. Suppose a 10-unit difference from placebo was deemed to be clinically useful. If the lower limit of the CI was greater than 10 then one would believe strongly that this was a clinically useful treatment. On the other hand, suppose that the *P* value for the null hypothesis of no treatment difference was small, but the lower limit of the CI was less than 10 and the upper limit was greater than 10. At this point one would believe that there is a treatment effect and that it could be a clinically important difference. Another example would be when the *P* value was small, but the upper limit of the CI was less than 10. At this point although

there is a statistical difference from placebo, the difference is smaller than what would be clinically useful. One last example would be when the *P* value was large, but the upper limit of the CI was greater than 10. In this case you have not proved there is a difference, but you have also not ruled out that a clinically useful difference exists. As can be seen, using these tools in unison provides a lot more knowledge than using either by itself.

There is a different type of statistical inference that can be done using Bayesian analyses.[5] These inferences overcome some of the logical traps of more traditional inference, but also have negatives associated with them, as well. I will not attempt to introduce this subject, but wanted to mention this alternative method to statistical inference. The Bayesian tools, if used properly, can quantify decision making to a further extent. With that said, the important point is that the use of statistical inference, if done correctly, helps science see a clearer picture of what their data are trying to tell them.

EXAMPLE

To demonstrate the improvement to inference by using mechanistic models, a small simple example follows. Data were simulated for a hypothetical clinical trial that enrolled 36 subjects. Each subject was randomly assigned to 1 of 6 doses (placebo, 1 mg, 5 mg, 10 mg, 40 mg, and 75 mg), so that each dose was assigned to 6 subjects. Drug concentration data were assumed to have a lognormal distribution with geometric mean of 0.239 × dose and coefficient of variation of 50%. Effect was assumed to have a lognormal distribution with coefficient of variation of 20% and geometric mean of $10 \times c/(6 + c)$, where c is the concentration.

Once the 36 observations were generated, an ANOVA was performed with the natural log of the effect as the dependent variable and dose as the independent variable. In addition, an E_{max} model was fit of the following form:

$$\text{Effect} = \left[E_0 + \frac{E_{\max} \times \text{concentration}}{EC_{50} + \text{concentration}}\right] \exp(\varepsilon) \qquad (1)$$

where ϵ is normally distributed with a mean of 0 and variance of σ^2.

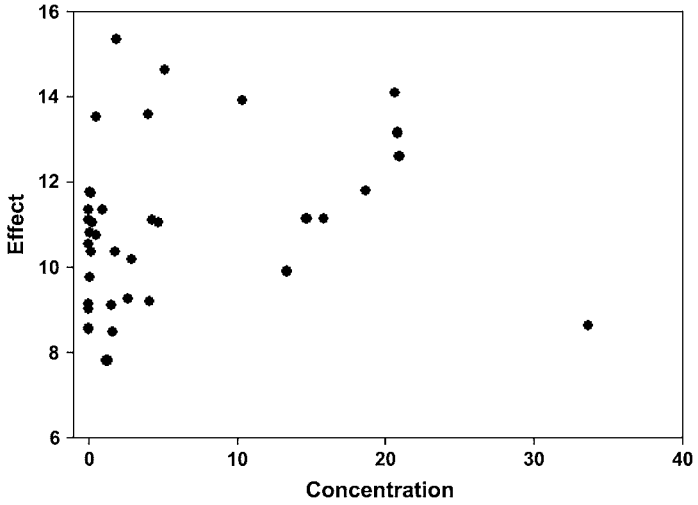

Figure 1. Scatter plot of simulated effect versus concentration

All analyses were performed with SAS version 8.02. (SAS Institute Inc, Cary, NC). The ANOVA used PROC MIXED. All pairwise comparisons to placebo were found using LSMEANS option. The E_{max} model was performed using PROC NLMIXED. To examine all pairwise comparisons to placebo, first, the geometric mean of concentration was found for each dose, then appropriate means and differences were found by estimating the effect at these concentrations using the ESTIMATE statement in PROC NLMIXED.[6]

Figure 1 displays the data used for the analyses. A subtle effect may be noticeable, but it is certainly less than dramatic.Table 1 shows the pairwise comparisons from the ANOVA.

It is assumed, in this case, that as the dose increases so does the effect. One of the irritating attributes of this display of information is that it does not recognize this assumption. Thus, we see no consistent pattern of increasing effect with increasing dose. In fact, we appear to be detecting a difference from placebo for the 40-mg dose, but not the 75-mg dose. Suppose that the drug needs to possess at least a 40% increase for it to be a viable candidate. The results for the 40-mg dose make this result seem tenable, but the results at 75 mg seem to show that it is not likely. As can be seen, the inference is not straightforward.

Use of mechanistic models takes care of this dilemma. First, for there to be sufficient evidence that the drug has an effect,

Table 1. ANOVA Comparisons

Treatment	Geometric Mean	Ratio to Placebo	95% Confidence Interval	*P* Value	Precision (UCL/LCL)
Placebo	9.89				
1 mg	10.9	1.10	(0.93, 1.30)	.25	1.40
5 mg	9.99	1.01	(0.86, 1.19)	.90	1.40
10 mg	10.7	1.09	(0.92, 1.28)	.33	1.40
40 mg	12.7	1.29	(1.09, 1.52)	.0041	1.40
75 mg	11.3	1.14	(0.97, 1.35)	.11	1.40

Table 2. Comparisons from E_{max} Model

Concentration (Median for Each Treatment)	Geometric Mean	Ratio to Placebo	95% Confidence Interval	*P* Value	Precision (UCL/LCL)
0 (Placebo)	10.3				
0.151 (1 mg)	10.4	1.01	(0.99, 1.02)	.51	1.03
0.993 (5 mg)	10.6	1.03	(0.96, 1.11)	.42	1.16
2.81 (10 mg)	11.0	1.06	(0.95, 1.20)	.28	1.26
8.35 (40 mg)	11.4	1.11	(0.98, 1.26)	.11	1.29
19.4 (75 mg)	11.7	1.13	(0.99, 1.30)	.08	1.32

the parameter of E_{max} must be shown to be greater than 0. For this example the estimate of E_{max} was 1.68 (SE = 1.11). The *P* value associated with the hypothesis that $E_{max} = 0$ is .14. There does not appear to be a great amount of evidence to suggest that the drug causes an effect. Table 2 shows the pairwise comparisons from the E_{max} model.

Notice that the inference about a 40% increase is pretty straightforward. It does not look like the drug can deliver a 40% treatment difference.

The last column of each of the 2 tables (UCL/LCL) is a measure of precision, which is the upper limit of the confidence interval divided by the lower limit. Notice that the mechanistic model is uniformly more precise than the inference from ANOVA. In fact, it can be shown that for an ANOVA model to have the same degree of precision as the E_{max} model (ratio of 1.32), one would need almost 50% more subjects. Thus, not only do mechanistic models provide more straightforward inference, but they also do it more efficiently. The 50% number comes from the following. The standard error for ANOVA is

$$SE = s\sqrt{\frac{1}{\eta_1} + \frac{1}{\eta_2}},$$

where s is the square root of the mean square error and n_i is the sample size of the ith subgroup. Since the sample sizes are equal in each group this reduces to

$$SE = s\sqrt{\frac{2}{\eta}}.$$

The width of the 90% confidence interval $w = 2 \cdot t_{0.1,6\cdot(n-1)} \cdot SE$. For the ANOVA, $w = \ln(1.4)$. For larger sample sizes, the t value changes little. Assuming no change in the t value, then one could find that the sample size needed to produce a width of ln(1.32), n_{new}, relative to the sample size that produced a width of ln(1.4), which is given by

$$\frac{n_{new}}{n_{old}} = \left(\frac{\ln(1.4)}{\ln(1.32)}\right)^2 = 1.47 \qquad (2)$$

One must recognize how important it is for the mechanistic model to be well specified. Potentially a large amount of bias is possible for a poorly specified mechanistic model. If we want to get these inferential advantages that will drive cost savings, then an individual who is accountable for the form of the model (the scientist) must exist.

Conclusion

The usefulness of statistical inference has been discussed in this paper. The ability to have more powerful inferences through the use of proper mechanistic models has also been demonstrated. Yet, sadly, even with these facts, the two are seldom combined together. If information were cheap, we could easily ignore these facts and go on relegating data interpretation separately to the pharmacometrician and statistician. Everyone could happily and ignorantly go about his or her business. The fact, however, is that information is not cheap. It is imperative that appropriate decisions be made at the smallest cost possible. Because the pharmaceutical industry is not as lucrative as it once was, we have to come up with cheaper ways to develop our molecules. Therefore, it is imperative that we get the full use out of our mechanistic modeling efforts, which will only be done when we break down the barriers that we have set up between the professions of statistics and pharmacometrics. It's time.

References

1. Salsburg D. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century.* New York: Henry Holt and Company; 2002.

2. Matthey S. p<.05—But is it clinically significant? Practical examples for clinicians. *Behav Change*. 1998;15:140-146.

3. Dixon P. The p-value fallacy and how to avoid it. *Can J Exp Psychol*. 2003;57:189-202.

4. Moyé LA. *Multiple Analyses in Clinical Trials.* New York: Springer-Verlag; 2003.

5. O'Hagan A, Luce BR. *A Primer on Bayesian Statistics in Health Economics and Outcomes Research.* Sheffield: Centre for Bayesian Statistics in Health Economics 2003; Available at: http://www.shef.ac.uk/~st1ao/pdf/primer.pdf. Accessed June 1, 2005.

6. *SAS/STAT User's Guide.* Version 8. Cary, NC: SAS Institute Inc; 1999.